**Stanford eCorner**

**Unanswered AI Questions**

**23-02-2024**

URL: https://stvp.stanford.edu/clips/unanswered-ai-questions

Daniela Amodei, president and co-founder of Anthropic, notes that fundamental questions that are still unanswered about how large-language models work, so her company publishes its research on the topic. Anthropic's long-term research focuses on understanding how neural networks work in the same way neuroscientists study the human brain.

## Transcript

- I spent a lot of time helping nonprofit organizations 00:00:06,450 understand the potential of AI, but also the risks.. And it's not always apparent that there's some challenges with these tools.. It's not always accurate.. - Yep.. 00:00:15,882 - Large language models can hallucinate.. 00:00:19,530 Depending on the data you provided, there could be bias.. There could be essentially people unfortunately putting private information into public models.. How do you see Anthropic's role, or your role, or our role as citizens in understanding what, how we need to show up to be a good partner to these tools? - Yeah, I think I have sort of a like short term answer, 00:00:44,040 which is like, what do we do with kind of what is available to us today? And then kind of a like more speculative, kind of long term answer.. So I think on this sort of, you know, short term front, I think there is a lot of just really interesting work being done around some of these fundamental questions, right? What does it mean to sort of use these models and understand what's happening inside them? And we try to publish our research about all of this, right? We don't have perfect information because these models are, even to the people that are training them, still like a little bit of a mystery, right? We know, hey, you put in data, you put in compute, you do some fancy magical algorithms and like magic, right? You have these really powerful tools, but all of the sort of details underneath are a little bit opaque still.. And so I think to the degree that we are kind of able to, whether it's with, you know, customers or lawmakers or individuals sort of explaining what we've done is kind of really a big part of the ethos of Anthropic..

In the longer term, I sort of particularly want to click into a research team at Anthropic in the area of mechanistic interpretability.. Which is an area that one of my other co-founders sort of, you know, pioneered, first at Google actually, and then at OpenAI and now at Anthropic.. And really the best way to think of mechanistic interpretability is almost as the equivalent of like neuroscience and what neuroscientists do to the human brain, what mechanistic interpretability experts do to neural networks.. And so really thinking about, you know, when models, when these sort of neural networks are producing outputs, we don't know what's happening.. Just like when humans are sort of thinking, "Oh, like I think this thing about this person, why do I think that?" If we could actually go in and say, what are the literal neurons that are firing that are causing the model to think this or do this? And are there combinations that are maybe problematic that are firing together? And so even if you can sort of train it out of the model at the end, it would be much better if we saw are there kind of problematic things or positive things happening in the model that we would want to adjust?..